Machine Learning & Prediction

Irene Yao

4/16/2018

Based on the previous data analysis (https://github.com/ireneyaoyao/Springboard/blob/master/Capstone/Statistical%20Analysis.pdf), we will use the following variables for prediction of outcome type.

- size
- intake_condition
- outcome_condition
- sex
- age_at_intake
- stage_at_outcome (age)
- days in shelter
- breed (is mix or not)

Machine Learning and Prediction

For this report's purpose, the prediction will be made around dogs, and the predicted value will be the outcome of each animal. I will use GBM for the prediction and the outcome will be a binary classification. The outcomes for the animals will be either "placed in a home" or "not placed in a home". "Adoption" and "Return to owner" will regarded as "placed in a home", and all others will be categorized into "not place in a home". To do so, I will add a column "placed" and the binary value for the column will be 1 or 0.

```
dogs$placed <- ifelse((dogs$outcome_type == "ADOPTION" | dogs$outcome_type == "RETURN TO OWNER"
), 1, 0)
```

Some of the columns have a class of "character" or "timediff". In order for the prediction model to work, update those columns to either "factors" or "numeric".

```
dogs$sex_clean <- as.factor(dogs$sex_clean)
dogs$stage_at_outcome <- as.factor(dogs$stage_at_outcome)
dogs$age_at_intake <- as.numeric(dogs$age_at_intake)
dogs$age_at_outcome <- as.numeric(dogs$age_at_outcome)
```

Prediction Using GBM

1. separate the dataframe into a training and a testing set. 80% of data will be in training set and the rest 20% will be in testing set.

```
n <- nrow(dogs)
n_train <- round(n * 0.8)
set.seed(123)
train_indices <- sample(1:n, n_train)
dog_train <- dogs[train_indices, ]
dog_test <- dogs[-train_indices, ]
```

2. create the GBM model

```
library(gbm)
set.seed(1)
dog_model_gbm <- gbm(formula = placed ~ size + intake_condition + outcome_condition + sex_clean
 + age_at_intake + stage_at_outcome + is_mix,
                             distribution = "bernoulli",
                             data = dog_train,
                             n.trees = 10000)
```

3. predict the outcomes of the test set

```
pred_gbm <- predict(object = dog_model_gbm,
                newdata = dog_test,
                n.trees = 10000,
                type = "response")
```

4. evaluate the model using test set AUC

```
library(Metrics)
auc <- auc(actual=dog_test$placed, predicted=pred_gbm)
print(paste0("Test set AUC: ", auc))
```

```
## [1] "Test set AUC: 0.873844537815126"
```